SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100304

Nouveau sens et évolution des domaines d'emploi : méthodologie pour l'acquisition lexicale

Reutenauer, Coralie

ATILF (CNRS & Université de Lorraine)
coralie.reutenauer@atilf.fr

1 Introduction

Le sens se construit à travers des interactions permanentes entre langue et discours. Les traitements automatiques disposent de ressources et d'outils de plus en plus développés pour simuler les interactions entre langue et discours. Des ressources électroniques en nombre croissant proposent des représentations de la langue, sous forme d'ontologies, de terminologies, ou encore de dictionnaires informatisés dont le format est plus ou moins largement enrichi par rapport à des versions papier préexistantes, voire même spécifique au format électronique (Béjoint, 2009). Ces ressources sont utilisées à différents niveaux pour analyser des corpus textuels représentatifs des discours : Rayson *et al.* (2004) procèdent à une annotation de corpus en parties du discours et en domaines issus du *Longman Lexicon of Contemporary English* dans une perspective d'analyse de contenu sémantique et de désambiguïsation, ou encore Victorri et Fuchs (1996) utilisent des réseaux de synonymie élaborés à partir de dictionnaires de synonymes pour calculer le sens d'un énoncé. L'interaction entre des dictionnaires électroniques et des ressources textuelles s'effectue alors dans un sens : les dictionnaires sont au service de la dynamique en corpus. À l'inverse, différents travaux utilisent les ressources textuelles pour enrichir des ontologies ou des ressources lexicographiques. Ainsi, Kilgarriff *et al.* (2004) proposent un outil d'extraction de « portraits de mots », le *Sketch Engine*, qui participe à la construction d'un dictionnaire, le *MacMillan English Dictionary for Advanced Learners*.

Les travaux évoqués gravitent autour de la question de la mise à jour de dictionnaires, mais celle-ci est loin d'être résolue, notamment en lexicographie française et pour le français standard. De fait, la mise à jour de ressources lexicographiques reste très largement manuelle et relève le plus souvent d'un « artisanat » (Rey, 2008). L'automatisation de l'acquisition lexicale reste un défi, particulièrement dans le cas de mots existants qui acquièrent de nouveaux sens, c'est-à-dire des néologies sémantiques. Le changement est plus difficile à détecter car il ne peut s'appuyer sur des critères formels tels que l'absence du mot de la ressource de référence. De plus, la construction du nouveau sens nécessite d'articuler le nouveau sens aux sens existants.

Dans une perspective d'automatisation de l'acquisition lexicale, il est nécessaire de prendre en compte quatre aspects : l'existence d'un nouveau sens en discours ; sa diffusion, et notamment sa stabilisation dans les discours ; la qualification du nouveau sens ; son articulation au sens existant. Certaines réalisations, notamment des plateformes de repérage et de stockage de néologies telles que celle de l'*Observatoire de Néologie* (Cabré *et al.*, 2003) ou la *Wortwarte* (Lemnitzer et Ule, 2011) contribuent à une facette du problème, à savoir la détection de néologies, mais elles laissent généralement de côté la question de la qualification du nouveau sens et de son articulation aux sens existants.

Deux types de travaux fournissent des éléments de réponse qui peuvent se combiner pour une analyse des différentes facettes du phénomène : des descriptions théoriques, essentiellement linguistiques, des mécanismes sémantiques en jeu dans les phénomènes de néologie, de diffusion sémantique ou d'émergence de nouveaux sens (Rastier et Valette, 2009 ; Sablayrolles, 2000) ; des travaux orientés vers une perspective applicative qui répertorient des indices de néologie ou de diffusion d'un sens (Picton, 2009). Nous nous positionnons à l'intersection de ces travaux pour proposer ce qui pourrait constituer la première étape de l'acquisition d'un nouveau sens : la caractérisation de l'évolution de sens en fonction de domaines.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100304

2 Problématique

Dans notre approche, la recherche d'informations sémantiques pertinentes repose sur un critère de variation sémantique marquée et répétée dans le temps. Si la variation sémantique est faible, l'intérêt de l'acquisition d'un nouveau sens est limité, puisque l'ancien sens fournit l'essentiel des éléments nécessaires à l'interprétation du nouveau sens. Sans répétition dans le temps, le nouveau sens n'est qu'un phénomène accidentel, qui ne correspond pas à un emploi qui vit effectivement dans les discours.

Les mécanismes en jeu dans la néologie sémantique reposent notamment sur des phénomènes génériques, qui se situent au niveau des domaines ou des thèmes. Rastier et Valette (2009) mettent en évidence différents mécanismes à l'origine de la néologie sémantique : des phénomènes de domanialisation, qui correspondent à l'émergence d'un sens propre à un nouveau domaine, comme le domaine de la gastronomie pour *moléculaire* (*cuisine moléculaire, spaghettis moléculaires*) ; des phénomènes de dédomanialisation, où le nouveau sens s'affranchit du ou des domaines spécifiques d'emploi, notamment pour évoluer vers un emploi général (par exemple, *mutualiser*, passé du domaine de l'assurance à des emplois généraux (Viprey et Schepens, 2010)). L'importance des informations génériques telles que les domaines dans le phénomène de néologie sémantique apparaît également chez Sablayrolles (2000). En effet, les procédés à l'origine de la néologie sémantique sont essentiellement des figures de style, notamment des métaphores ou des métonymies. Or celles-ci reposent sur des transgressions qui se situent à un niveau générique, que ce soit entre domaines ou champs sémantiques (Rastier, 1987). Les informations génériques telles que les domaines ne fournissent pas une information exhaustive sur le changement de sens, mais elles jouent un rôle fondamental. De ce fait, dans nos expériences, le changement de sens sera décrit exclusivement à l'aide de domaines.

Par ailleurs, divers travaux ont mis au jour des indices adaptés pour appréhender la néologie sémantique dans un contexte applicatif. Les indices de néologie sémantique peuvent être exploités sous deux angles et contribuer principalement à la détection d'un nouveau sens (indices tels que les guillemets ; *cf.* Bauer et Renouf (2000)) ou à sa qualification. Les indices qui contribuent à la qualification du nouveau sens sont notamment les cooccurrents lexicaux sur l'axe syntagmatique et, sur l'axe paradigmatique, les concurrents issus du foisonnement néologique (par exemple, le triplet {*manageur, décideur, directeur*} (Cusin-Berche, 2003)). Un autre indice est celui des empreintes de fréquence (Picton, 2009) qui correspondent à l'évolution au cours du temps de la fréquence d'une unité lexicale donnée. Les empreintes de fréquence sont *a priori* plus adaptées pour participer à la détection d'un nouveau sens qu'à sa qualification, mais leur étude relativement à un espace défini par une structure sémantique les rend pertinentes pour qualifier le sens, comme nous le verrons en 3.1. L'intérêt des empreintes de fréquence est qu'elles accordent à la diffusion et au paramètre temps un rôle central. C'est cet indice que nous retenons pour la suite des développements, avec un objectif de qualification du nouveau sens.

Nous abordons donc l'émergence d'un nouveau sens comme un processus qui se construit en contexte et qui se caractérise par une double diffusion : dans le temps et dans des domaines. La détection d'un nouveau sens n'est pas étrangère à notre approche, mais la qualification du nouveau sens est notre objectif principal. Nous cherchons à y parvenir à partir d'indices mesurables et porteurs d'information sémantique spécifique de l'unité lexicale ciblée..

3 Indices pour analyser la diffusion par domaine d'un nouveau sens

Nous proposons d'aborder la diffusion du nouveau sens en jouant sur deux paramètres : le temps et les domaines d'emploi. Pour cela, nous nous appuyons sur l'indice de diffusion que sont les empreintes de fréquence, que nous adaptons à notre approche paramétrée et que nous articulons à un cadre statistique généralement utilisé pour l'analyse des cooccurrences.

3.1 Empreintes de fréquence par domaine

Les empreintes de fréquence, c'est-à-dire l'évolution dans le temps de la fréquence d'une unité lexicale, peuvent être utilisées pour mettre en évidence la diffusion d'une unité lexicale dans les discours. Dans notre cadre, nous cherchons à observer l'apparition de nouvelles lexies et leur implantation dans les discours. Nous faisons l'hypothèse qu'une néologie de sens peut s'accompagner d'un accroissement de fréquence, à même de représenter la diffusion d'un nouvel emploi dans les discours. Cet accroissement de fréquence peut se produire lorsque le sens de la lexie évolue d'un emploi propre à un domaine particulier vers un emploi plus général. C'est notamment le cas de *mutualiser*, initialement associé au domaine de l'assurance puis employé dans un cadre plus général à partir de la fin des années 90 et le début des années 2000 (Viprey et Schepens, 2010).

Les informations purement quantitatives délivrées par les empreintes de fréquence doivent toutefois être manipulées avec précaution. En effet, l'existence d'un accroissement de fréquence ne garantit pas qu'un nouveau sens s'implante. Tous les néologismes de sens ne présentent pas nécessairement un accroissement global d'emplois. Un nouvel emploi peut rester mineur par rapport aux emplois préexistants et n'influer que faiblement sur les variations de fréquence, ou un nouveau sens peut remplacer un sens préexistant, donc, en l'absence de désambiguïsation, la fréquence de l'unité lexicale peut stagner, voire diminuer.

L'apport des empreintes de fréquence se limite à la détection d'un nouveau sens tant que leur analyse n'est pas étayée par des précisions supplémentaires sur leur environnement sémantique, c'est-à-dire tant qu'elles sont découplées d'un profilage sémantique des discours. Mais si elles sont corrélées à des profilages sémantiques des discours, par exemple à des profilages thématiques ou domaniaux, elles peuvent contribuer à la qualification du nouveau sens en plus de sa détection.

3.2 Indices statistiques et découpages de l'espace textuel

Pour les empreintes de fréquence telles que nous les avons présentées, l'observable est la cible lexicale. Cependant, avec les empreintes de fréquences thématiques, il est possible d'adopter plusieurs perspectives et de se focaliser soit sur la cible, soit sur le paramètre d'étude que sont les domaines. Autrement dit, il est possible d'avoir une approche duale, qui participe soit à la détection lorsque l'attention se focalise sur la cible, soit à la caractérisation lorsque l'attention se focalise sur les domaines : on étudie la diffusion de la cible x dans le domaine y ou l'émergence du domaine y au voisinage de x.

Dans le second cas, lorsqu'on se focalise sur les domaines, la perspective rejoint celle des approches cooccurentielles, à cela près que ce ne sont pas des unités lexicales mais des domaines qui sont étudiés : l'analyse porte sur la distribution d'entités linguistiques au voisinage de la cible lexicale par rapport au reste du corpus. L'approche cooccurrentielle respecte deux contraintes particulières lorsqu'elle est articulée aux empreintes de fréquence thématiques : d'abord, le niveau d'observation est supra-lexical au lieu de lexical, car les observables sont les domaines, qui apportent de l'information sémantique plus générale et de granularité moins fine que les unités lexicales ; ensuite, le paramètre temps fait partie intégrante de l'analyse.

Selon les hypothèses de distributionnalité de Harris (1968), des entités cooccurrentes, c'est-à-dire partageant les mêmes contextes ou encore qui présentent les mêmes distributions dans l'espace associé aux ressources linguistiques, sont supposées de sens proche. Le parallèle avec les approches cooccurrentielles renforce l'argument selon lequel les empreintes de fréquence thématiques participent à la qualification du sens.

Les méthodes communément utilisées en lexicométrie pour identifier les cooccurrents saillants au voisinage de la cible lexicale et pour extraire des unités en affinité sémantique avec celle-ci s'appuient sur le calcul d'indices statistiques tels que le χ^2, l'information mutuelle ou les spécificités. Ces indices dépendent généralement d'un découpage de l'espace textuel en un sous-corpus d'étude et son complémentaire, ainsi que de 4 paramètres : 1) la taille totale du corpus ; 2) la taille du sous-corpus, *i.e.* le voisinage de la cible ; 3) le nombre total d'occurrences de l'unité observée dans le corpus ; 4) son nombre

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100304

d'occurrences dans le sous-corpus. Autrement dit, les fréquences, qui constituent l'unité de décompte à la base des empreintes de fréquence, sont intégrées dans les indices statistiques et le résultat du calcul peut se voir comme des fréquences réajustées en fonction de paramètres supplémentaires. Dans notre cadre, nous fondons nos analyses sur de tels indices. Plus précisément, nous utilisons les spécificités de Lafon (1984) comme indice statistique. Dans nos expériences, le calcul est réalisé par un module de la plateforme *Semy* (Grzesitchak, 2008)[1]. Il est similaire à celui du logiciel *Lexico3* (Salem *et al.,* 2003). Les valeurs retournées sont des entiers positifs en cas de surreprésentation, négatifs sinon. Plus la spécificité est grande en valeur absolue, plus la saillance de l'unité est considérée comme significative.

Le calcul des spécificités repose sur une partition du corpus en deux sous-ensembles : un sous-corpus et son complémentaire. Pour analyser le sens d'une cible lexicale, le sous-corpus correspond au voisinage de la cible. Dans notre cadre, la structure du corpus est plus complexe : les domaines et les périodes de temps sont des paramètres qui permettent des découpages multiples. Il est ainsi possible de définir un voisinage de la cible propre à tout couple (domaine, période). Deux modes de calcul des spécificités permettent d'appréhender cette structure de l'espace textuel. Le premier est effectué à domaine fixé, période par période, ce qui permet d'obtenir une suite de spécificités caractéristiques d'un domaine donné, indépendamment des autres domaines. Le deuxième est destiné à refléter le poids relatif des domaines, autrement dit la ventilation des domaines, et la façon dont évolue cette ventilation dans le temps. Les spécificités sont calculées à période fixée, et ce pour chacune des périodes, ce qui permet d'avoir un aperçu de la façon dont se répartissent successivement les domaines les uns par rapport aux autres. Les deux modes de calcul correspondent à deux axes d'analyse complémentaires. Le premier permet d'appréhender la diffusion au sein d'un domaine, c'est-à-dire l'accroissement des emplois au cours du temps pour le domaine considéré. Le deuxième permet d'observer l'évolution de la ventilation, c'est-à-dire la façon dont se reconfigure le sens au cours du temps à travers l'évolution de l'équilibre relatif entre domaines.

Les deux points de vue doivent contribuer à l'acquisition d'informations qualitatives nuancées, destinées à mettre en évidence une variété de sous-phénomènes : l'existence d'une diffusion domaniale simple ou multiple, la reconfiguration du sens, la façon dont progresse la diffusion et les affinités ou oppositions entre domaines, en particulier entre anciens et nouveaux domaines d'emploi. Chacun de ces points fera l'objet d'un développement, étayé par des expériences en corpus.

4 Ressources textuelles, ressources lexicographiques et cibles lexicales

La diffusion d'un nouveau sens est abordée sous l'angle de la linguistique de corpus (4.1) à partir de ressources textuelles où sont observées différentes cibles lexicales (4.2).

4.1 Un corpus multidomaines

Le corpus d'étude est issu de la base de données d'actualité internationale *Factiva*. Cette base de données, constituée de plus de 10 000 sources, notamment journalistiques (presse nationale et internationale), a été restreinte à une sélection de journaux francophones et de thématiques ainsi qu'à une période donnée. Les sources sélectionnées sont les journaux *Libération, Le Figaro, L'Humanité, Ouest-France, La Tribune, Les Echos* et *L'Expansion*. Les thématiques sont les suivantes : informations économiques (ECO), arts et spectacles (ART), environnement (ENV), mode de vie (MOD), politique / relations internationales (POL), santé (SAN), science et technologie (SCI), société / communauté / travail (SOC). Ces thématiques

[1] La plateforme Semy est une plateforme d'annotation sémantique qui enrichit les corpus en informations extraites du *Trésor de la Langue Française informatisé*. Elle est dotée de modules lexicométriques, permettant d'effectuer des calculs statistiques, dont le calcul des spécificités. Elle fait partie de ressources internes encore en développement et n'est actuellement pas en libre diffusion.

DOI 10.1051/shsconf/20120100304

sont choisies parmi les sujets intégrés au moteur de recherche de Factiva. Nous les qualifierons par la suite de domaines. Le corpus s'étend de 2004 à 2010, il est subdivisé en tranches annuelles.

Le corpus compte 1,2 million d'articles. Sa restriction aux deux années 2004 et 2010, utilisée pour les expériences des sous-sections 5.1 et 5.2, est d'environ 300 000 articles. L'unité textuelle est l'article et elle sert également d'unité de décompte des occurrences : la taille d'un sous-corpus ou encore les occurrences d'une cible lexicale seront évaluées en nombre d'articles.

4.2 Des cibles lexicales à profils variés

Les cibles présélectionnées sont *toxique*, *dangereux* ; *délétère ; tsunami* ; *tempête ; raz-de-marée* ; *moléculaire* ; *tablette*. Elles se répartissent en deux catégories :

- **des cibles pressenties ou identifiées comme néologiques.** La première est l'adjectif *toxique* en contexte de crise financière, qui qualifie des produits ou instruments financiers à l'origine de la crise financière de 2008 (les *actifs toxiques*, étroitement rattachés aux subprimes). La deuxième cible est *tsunami*, employé dans un sens métaphorique à l'instar de *tempête* ou de *séisme* pour qualifier des événements désastreux ou catastrophiques. Notre hypothèse est que la diffusion des emplois métaphoriques est déclenchée par la catastrophe naturelle qui a frappé l'Asie du Sud-est le 26 décembre 2004[2]. La 3e cible est *moléculaire*, qui connaît de nouveaux emplois dans le domaine de la gastronomie (*cuisine moléculaire*). Ceux-ci renvoient à un jeu sur les textures, les couleurs et les formes en cuisine créant un effet artistique et évoquant une pratique régie par des principes scientifiques. La dernière cible est *tablette*. Les nouveaux emplois sont liés à une innovation technologique, la *tablette numérique*, qui connaît matériellement une diffusion de grande ampleur dans la société en 2010[3] et, de façon corrélée sur le plan linguistique, une diffusion massive dans les discours.
- **des cibles non néologiques, utilisées comme témoin.** Il s'agit de *raz-de-marée, tempête, délétère et dangereux. Tempête* et *raz-de-marée* appartiennent au paradigme des phénomènes météorologiques, comme *tsunami*. Ils peuvent être employés au sens figuré. Les emplois métaphoriques de *tempête* sont particulièrement répandus[4]. Ils permettent de qualifier des bouleversements importants. De ce fait, les domaines d'emploi sont susceptibles de varier en fonction des événements dominants de l'actualité et de présenter une volatilité d'un domaine à l'autre qui pourrait être le propre de tels emplois métaphoriques. *Dangereux* et *délétère* ont été choisis pour leux proximité sémantique avec *toxique*. Le premier est d'usage très répandu, l'autre est plus rare. Ils se distinguent par leur importance dans l'usage, par rapport à laquelle *toxique* se situe de façon intermédiaire.

5 Construction du profil domanial des cibles lexicales

Pour établir le profil d'une cible lexicale en corpus en fonction des domaines, nous procédons en trois étapes : vérification de l'existence d'une diffusion des emplois, simple ou commune à plusieurs domaines ; identification des domaines émergents pour les cibles retenues ; analyse de la progression de la diffusion dans le temps relativement aux différents domaines.

5.1 Existence d'une diffusion dans un ou plusieurs domaines

La première étape consiste à valider ou éliminer les cibles lexicales présélectionnées selon qu'elles présentent ou non une diffusion et à déterminer si la diffusion est propre à un domaine ou commune à

[2] Dans nos expériences, les résultats sont calculés par tranches de temps annuelles et l'année 2004 est utilisée comme année de référence pour les anciens emplois. Pour éviter un biais dû au pic événementiel des derniers jours de 2004, nous avons éliminé les documents datant du 25 au 31 décembre 2004 pour l'étude de *tsunami*.

[3] La tablette graphique existe depuis plusieurs décennies, mais ce produit a connu un bond technologique très récemment, à l'origine du changement. L'amorce du changement a eu lieu en 2007 avec le lancement de l'iPhone, proche des tablettes en termes de fonctionnalités et de technologie, et le véritable tournant a eu lieu en 2010 avec la mise sur le marché de l'iPad.

[4] A titre d'illustration, lorsque Perlerin (2004) évoque le paradigme de la météorologie boursière, une métaphore conceptuelle, il cite fréquemment *tempête*, *raz-de-marée* n'apparaît qu'une fois au détour d'un exemple et *tsunami* est absent.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100304

plusieurs domaines. Une précision complémentaire est apportée dans un second temps, pour déterminer si la diffusion est nouvelle ou si elle amplifie un usage déjà répandu.

À ce stade, nous recherchons une caractérisation relativement tranchée. Le tri des cibles et la qualification de la diffusion sont obtenus par une approche contrastive de deux périodes de temps disjointes, l'une ancienne, l'autre récente. Dans notre corpus, les périodes choisies sont les années 2004 et 2010. On utilise un critère d'***accroissement de la diffusion,*** évalué à partir des spécificités calculées à domaine fixé pour le sous-corpus de l'année 2010 par rapport au corpus des deux années réunies (*cf.* figure 1).

	ART	ECO	ENV	MOD	POL	SAN	SCI	SOC
dangereux	-14	-1	-4	-3	-6	-2	-1	2
délétère	-2	1	0	1	4	4	0	3
moléculaire	0	0	0	-2	-1	1	-3	0
raz-de-marée	-2	-2	-2	-6	-10	-2	1	-4
tablette	**14**	**9**	3	2	**8**	0	**5**	3
tempête	-3	**27**	1	-1	**22**	0	1	**11**
toxique	3	**23**	-4	1	4	-2	0	2
tsunami	**8**	**5**	2	**5**	**33**	2	1	**13**

Figure 1. Spécificités dans chaque domaine en 2010 relativement à 2004 calculées à domaine fixé pour l'ensemble des cibles lexicales (en gras : spécificités supérieures au seuil de 5).

Les cibles valident le critère de diffusion s'il y a accroissement significatif de la présence de la cible en deuxième période, dans au moins un domaine. L'accroissement est considéré comme significatif s'il franchit un seuil de spécificité, que nous fixons ici à 5[5]. Après application des seuils de spécificité, les cibles *tablette*, *tempête, toxique* et *tsunami* sont retenues, tandis que les cibles *dangereux, délétère*, *moléculaire* et *raz-de-marée* sont éliminées.

Pour les cibles retenues, nous distinguons deux types de diffusion selon qu'un ou plusieurs domaines émergent. L'intérêt est de dissocier deux phénomènes linguistiques, correspondant soit à l'ajout d'un nouveau domaine d'emploi, soit à la généralisation d'un emploi jusqu'alors propre à un domaine donné. Cette information est utile au niveau de la mise à jour du sens codé : dans un cas, il y aura ajout d'une étiquette de domaine et d'une nouvelle définition spécifique à ce domaine ; dans l'autre, la définition sera générale, sans étiquette de domaine. Dans le premier cas, l'accroissement permet d'isoler un domaine particulier. C'est le cas de *toxique*, où seul le domaine ECO présente un accroissement significatif, nettement supérieur à celui des autres domaines. Dans le second cas, l'accroissement est à peu près similaire dans plusieurs domaines. Cette configuration témoigne d'une diversification des domaines d'emploi et potentiellement de la diffusion d'un emploi métaphorique. Ce cas se présente pour *tsunami* (5 domaines émergents), *tablette* (4 domaines émergents) et *tempête* (3 domaines émergents).

Le critère d'accroissement de diffusion permet donc d'effectuer un premier tri entre les cibles, globalement conforme aux connaissances préalables sur les cibles et leur caractère néologique, à deux exceptions près. Trois des cibles pressenties comme néologiques, à savoir *toxique*, *tsunami* et *tablette* sont identifiées comme néologiques. En revanche, *moléculaire* n'est pas retenu, bien qu'il ait été pressenti comme néologique. Son emploi en gastronomie relève d'un champ trop spécialisé, pas suffisamment représenté par rapport aux autres domaines. De plus, la diffusion n'est pas suffisamment importante pour que les seuils de significativité soient franchis. Les cibles témoins non néologiques, à savoir *dangereux, délétère* et *raz-de-marée*, sont éliminées par la procédure, conformément aux attentes. En revanche, *tempête* n'a pas été éliminé. Une explication possible est que l'accroissement est dû aux emplois figurés de *tempête* et que les domaines saillants dépendent d'événements majeurs de l'actualité. Les observations

[5] Pour le choix du seuil, nous avons repris le seuil de spécificité proposé par défaut par le logiciel Lexico3 (Salem *et al.*, 2003), qui utilise le même mode de calcul des spécificités.

réalisées jusque-là ne sont pas suffisantes pour trancher entre de nouveaux emplois et l'expression d'un sens figuré préexistant. Les étapes ultérieures serviront à la question (à partir du 5.3).

5.2 Affectation d'un ou plusieurs domaines à chaque cible

L'objectif de cette étape est de sélectionner des domaines saillants pour les cibles retenues.

Les données précédentes peuvent être réutilisées pour associer à chaque cible un ou des domaines où se diffuse vraisemblablement le nouveau sens. Jusque-là, l'information a servi à valider certaines propriétés ou à quantifier le phénomène de diffusion plutôt qu'à qualifier le nouveau sens par des éléments nouveaux. L'attention était focalisée sur des informations quantitatives, telles que des franchissements de seuils ou le nombre de domaines émergents. Les informations recherchées ici sont qualitatives. Les sorties du traitement sont des étiquettes de domaines susceptibles de qualifier le nouveau sens.

Les domaines émergents peuvent être repérés de deux façons : (1) à travers un accroissement significatif de la présence de la cible au cours du temps et au sein du domaine considéré (étude à domaine fixé) ; (2) à travers l'émergence d'un nouveau domaine dans la tête de liste des domaines d'emplois (comparaison des ventilations de domaines). Pour combiner ces deux axes d'analyse, le calcul des spécificités à domaine fixé déjà réalisé est complété par des calculs de spécificités à période fixée, qui précisent la ventilation des domaines, c'est-à-dire leur importance les uns par rapport aux autres. Les résultats selon chaque axe d'analyse sont présentés en figure 3, avec un tri par spécificité décroissante des domaines pour chaque période et chaque cible.

tablette		*tempête*		*toxique*		*tsunami*	
2004	2010	2004	2010	2004	2010	2004	2010
MOD (13)	MOD (14)	ART (17)	POL (9)	ENV (125)	ENV (42)	ART (2)	POL (3)
SAN (3)	SCI (4)	ENV (16)	ENV (8)	SAN (60)	SAN (21)	SCI (2)	ENV (2)
ART (2)	ART (3)	SCI (2)	ECO (3)	SCI (16)	SCI (9)	ENV (1)	SCI (2)
SCI (1)	ECO (1)	MOD (1)	SCI (1)	MOD (-2)	ECO (4)	MOD (1)	SOC (2)
ENV (-2)	ENV (0)	POL (-3)	SOC (-1)	SOC (-3)	SOC (1)	SAN (0)	SAN (1)
ECO (-3)	SAN (-2)	SAN (-3)	SAN (-6)	POL (-17)	MOD (-4)	ECO (-1)	ECO (-2)
SOC (-3)	SOC (-5)	SOC (-8)	MOD (-7)	ECO (-24)	POL (-7)	POL (-1)	ART (-3)
POL (-5)	POL (-10)	ECO (-16)	ART (-9)	ART (-46)	ART (-24)	SOC (-2)	MOD (-3)

Figure 3. Ventilation des domaines pour chaque cible en 2004 et 2010, triés par spécificité décroissante

Les domaines émergents sont sélectionnés à partir des spécificités calculées à domaine fixé. Les ventilations sont utilisées pour préciser l'analyse : l'émergence d'un domaine n'est plus considérée isolément, mais relativement aux autres domaines (*cf.* figure 4).

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100304

	Domaines avec un accroissement marqué en 2010 (domaine fixé)	Évolution de la ventilation des domaines
toxique	ECO	ECO (Informations économiques) s'impose dans la ventilation : il est le deuxième domaine le plus sous-représenté en 2004 et il prend le pas sur trois autres domaines en 2010.
tsunami	POL, SOC, ART, ECO, MOD	Les ventilations sont faiblement contrastées en 2004 et en 2010, mais la configuration évolue entre les deux années. ART (Arts et spectacles) et MOD (Mode de vie) s'effacent devant les autres domaines. À l'inverse, SOC (Société / Communauté / Travail) et plus particulièrement POL (Politique / Relations internationales) s'imposent.
tempête	ECO, POL, SOC	Il y a évolution d'une ventilation contrastée vers une ventilation plus faiblement contrastée. ECO (Informations économiques) et POL (Politique / Relations internationales) se renforcent par rapport aux autres domaines, tandis que SOC (Société / Communauté / Travail) évolue de façon moins remarquable.
tablette	ART, ECO, POL, SCI	La ventilation est moyennement contrastée en 2004 et en 2010 et sa configuration reste approximativement la même. Deux domaines gagnent des rangs dans la ventilation, à savoir ECO (Inforamtions économiques) et SCI (Science et technologie).

Figure 4. Domaines associés à chaque cible à partir des spécificités calculées à domaine fixé et analyse complémentaire à l'aide des ventilations

Par recoupement des figures 3 et 4, le profil des domaines émergents varie selon les cibles. Pour *toxique*, ECO se distingue nettement des autres domaines. Il s'impose en terme de diffusion et de poids relatif par rapport aux autres domaines. Pour *tsunami*, la diffusion concerne les domaines rattachés aux sciences humaines et sociales, moins les domaines des sciences du vivant ou les autres sciences dures. Pour *tempête,* la diffusion concerne trois domaines, dont deux principalement, ECO et POL. La diffusion s'accompagne d'une reconfiguration de l'importance relative des domaines, même si les contrastes sont peu marqués. Pour *tablette*, la diffusion affecte plusieurs domaines, mais elle semble plus spécifique à deux domaines, SCI et ECO, qui connaissent un accroissement à domaine fixé et qui supplantent au cours du temps d'autres domaines dans la ventilation.

De façon générale, il y a convergence entre les deux axes d'analyse, à domaine fixé et selon la ventilation des domaines. L'intérêt est qu'ils fournissent des éclairages complémentaires, notamment lors d'une diffusion dans plusieurs domaines. Les exemples de *tsunami* et *tablette* sont particulièrement frappants : *tablette* se diffuse dans certains domaines en particulier, mais la ventilation des domaines reste presque la même en 2004 et 2010 et les contrastes demeurent ; *tsunami* se diffuse dans plusieurs domaines, mais la ventilation évolue vers une hiérarchie assez différente et l'échelle de spécificités est tassée : l'importance de ses emplois entre les différents domaines reste relativement homogène.

5.3 Analyse de la régularité de la diffusion des domaines

Aux étapes précédentes, les résultats sur les domaines ont été obtenus par comparaison de deux périodes de temps disjointes. Il est ainsi possible d'identifier un contraste, mais cela ne renseigne pas sur la progression de la diffusion. Celle-ci peut avoir des profils variés : diffusion linéairement croissante, pic de diffusion puis recul progressif (phénomène de mode ou événement marquant par exemple), croissance irrégulière, etc. Avec deux périodes seulement, on accroît le risque de sélectionner à tort certaines cibles candidates à un nouveau sens (par exemple, si la deuxième période correspond à un pic ponctuel) ou d'écarter des candidats valables (par exemple, si la deuxième période correspond à un creux dans une évolution irrégulière). La régularité de la diffusion n'est pas utilisée pour effectuer un nouveau tri des candidats mais pour nuancer l'analyse. Le contraste de deux périodes sert de critère initial pour l'allocation de signifié, même s'il n'exploite pas l'information liée à la continuité du processus. La

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100304

régularité de la diffusion est utilisée à titre complémentaire pour distinguer les candidats selon leur profil d'évolution. Dans nos expériences, le corpus d'étude est découpé en plusieurs périodes consécutives, sans discontinuité. L'évolution des domaines de 2004 à 2010 pour les cibles *toxique*, *tsunami, tempête* et *tablette* est détaillée ci-dessous. Pour chaque cible, deux graphiques d'évolution sont présentés et commentés : (1) les courbes d'évolution des spécificités à domaine fixé ; (2) les évolutions des ventilations, obtenues par calcul de spécificités sur l'ensemble des domaines à période fixée.

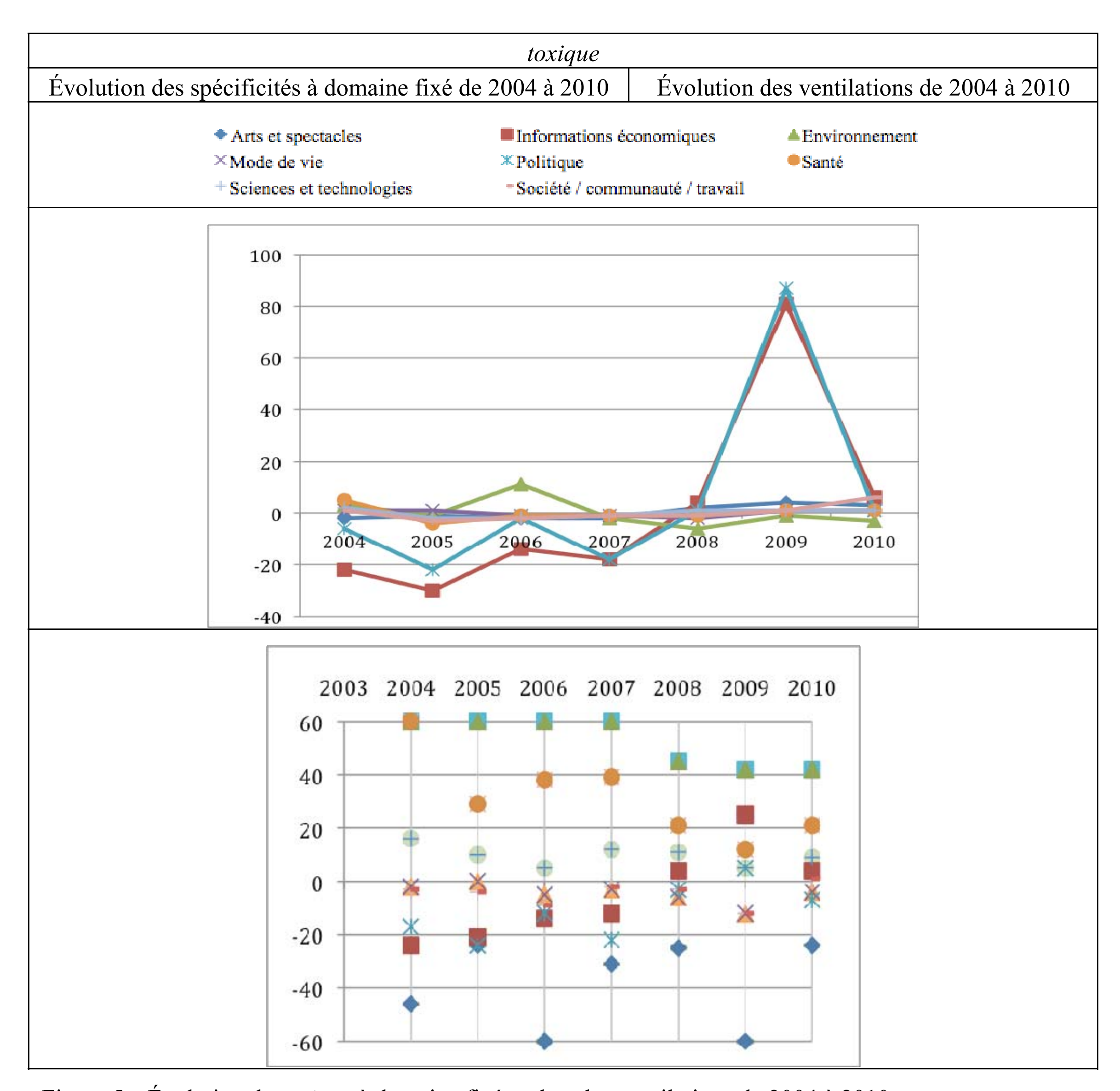


Figure 5a. Évolution de *toxique* à domaine fixé et dans les ventilations de 2004 à 2010.

Analyse de la cible *toxique* (figure 5a). À domaine fixé, pour la plupart des domaines, il n'y a pas de variation marquée au cours du temps. En revanche, en ECO et POL, il y a un contraste entre les périodes de 2004 à 2007, où les emplois sont sous-représentés, et 2008 à 2010, où ils sont surreprésentés, avec un pic en 2009. Dans la ventilation, un domaine évolue nettement : ECO se renforce au cours du temps. Il est l'un des domaines les plus fortement sous-représentés de 2004 à 2007 (6e et 7e rangs, avec des spécificités négatives), puis il rejoint les principaux domaines d'emploi. En revanche, POL reste parmi les derniers domaines d'emploi quelle que soit la période. L'évolution est relativement stable pour les autres

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100304

domaines. Le recoupement des deux axes d'observation montre une émergence nette d'un domaine particulier, ECO, conformément à l'analyse des deux années 2004 et 2010. L'accroissement est progressif, avec un basculement en 2008 qui se maintient en 2009 et 2010. Ceci va dans le sens d'une domanialisation de *toxique* en économie.

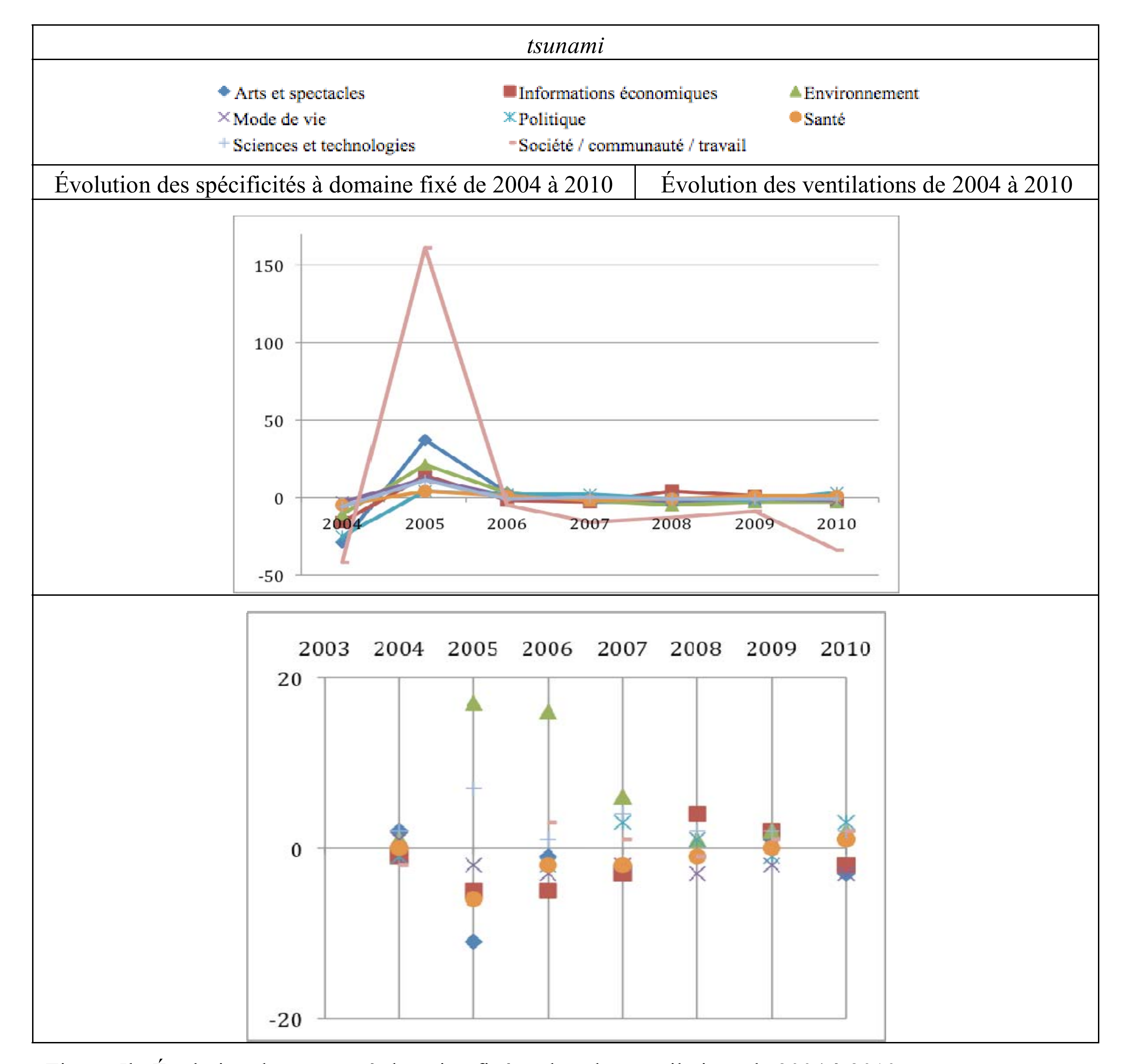


Figure 5b. Évolution de *tsunami* à domaine fixé et dans les ventilations de 2004 à 2010

Analyse de la cible *tsunami* (figure 5b). À domaine fixé, on constate que, quel que soit le domaine, il y a un pic en 2005 et la période où les spécificités sont les plus fortement négatives est 2004 : les emplois à cette période sont sous-représentés par rapport à ceux des autres périodes, autrement dit, il y a accroissement des emplois par rapport à 2004 dans toutes les périodes ultérieures, même si cette évolution est faible pour SAN.

Dans l'évolution des ventilations, quelques domaines sont stables : les domaines ENV (Environnement) et SCI (Science et technologie) ont un positionnement à peu près stable parmi les domaines les plus surreprésentés, toujours compris entre le 1[er] et le 3[e] rang, tandis que SAN est généralement sous-représenté. Le positionnement des autres domaines est en revanche instable, avec des évolutions

irrégulières et marquées d'une année sur l'autre. L'évolution période par période confirme l'émergence d'un nouveau sens, mais leur analyse amène à reconsidérer la liste des domaines émergents. Le changement est vraisemblablement déclenché par un pic événementiel et il s'accompagne d'un accroissement des emplois dans l'ensemble des domaines. L'évolution de la ventilation des domaines peut s'interpréter comme la diffusion d'un emploi métaphorique qui se généralise car, après le pic, les déséquilibres entre domaines sont de moins en moins nets : tous les domaines deviennent domaines d'emploi. De plus, la configuration n'est pas stable, avec des variations marquées d'une année sur l'autre, comme si la position était plus un phénomène ponctuel, dû à l'actualité, qu'une tendance intrinsèquement liée à la cible *tsunami*.

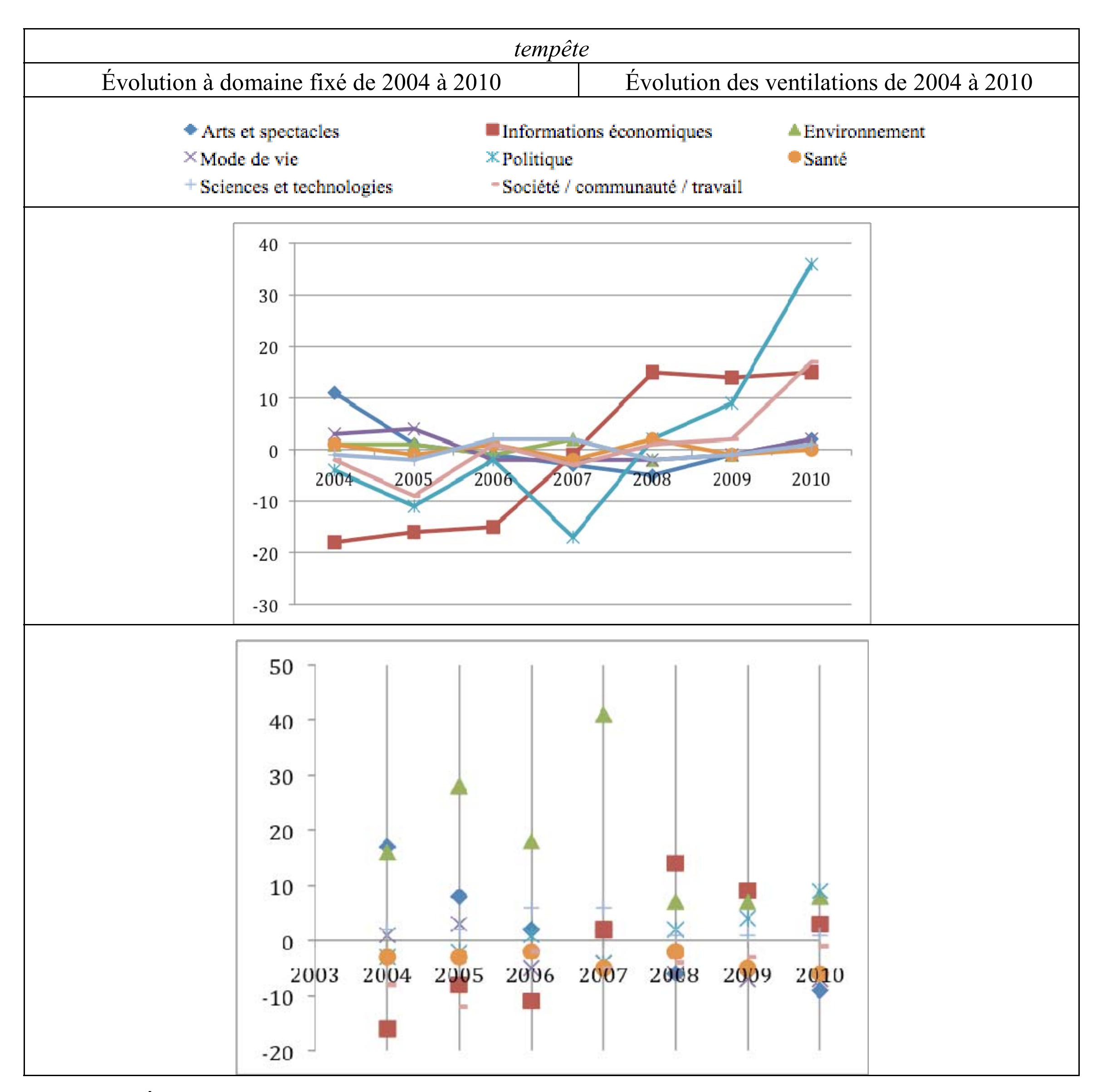


Figure 5c. Évolution de *tempête* à domaine fixé et dans les ventilations de 2004 à 2010

Analyse de la cible *tempête* (figure 5c). À domaine fixé, les évolutions sont très variables. ECO (Informations économiques), POL (Politique) et SOC (Société) présentent globalement un accroissement, mais qui n'a pas la même allure et avec des irrégularités. Il n'apparaît pas de cohérence globale, chaque

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100304

profil diffère ce qui donne une impression d'anarchie au niveau de l'évolution de l'ensemble des domaines.

Au niveau de la ventilation, tout comme pour *tsunami*, il y a stabilité pour ENV, SCI et SAN, tandis que les autres domaines connaissent des variations nettes dans la ventilation. ECO (Informations économiques) et POL (Politique) évoluent de domaines fortement sous-représentés vers les domaines les plus fortement surreprésentés, avec un basculement respectif en 2008 et 2007. SOC (Société) reste parmi les domaines aux spécificités les plus faibles. À l'inverse, ART (Arts et spectacles) s'efface de façon spectaculaire (passage progressif du premier au dernier rang).

Par recoupement des deux axes d'analyse, il y a apparemment émergence progressive de deux domaines, ECO et POL, qui apparaissaient déjà ans les résultats sur deux périodes disjointes. Cependant, le manque de stabilité dans la ventilation et les irrégularités d'évolution, similaires à celles de *tsunami,* pourraient être l'indice d'un comportement caractéristique des emplois métaphoriques.

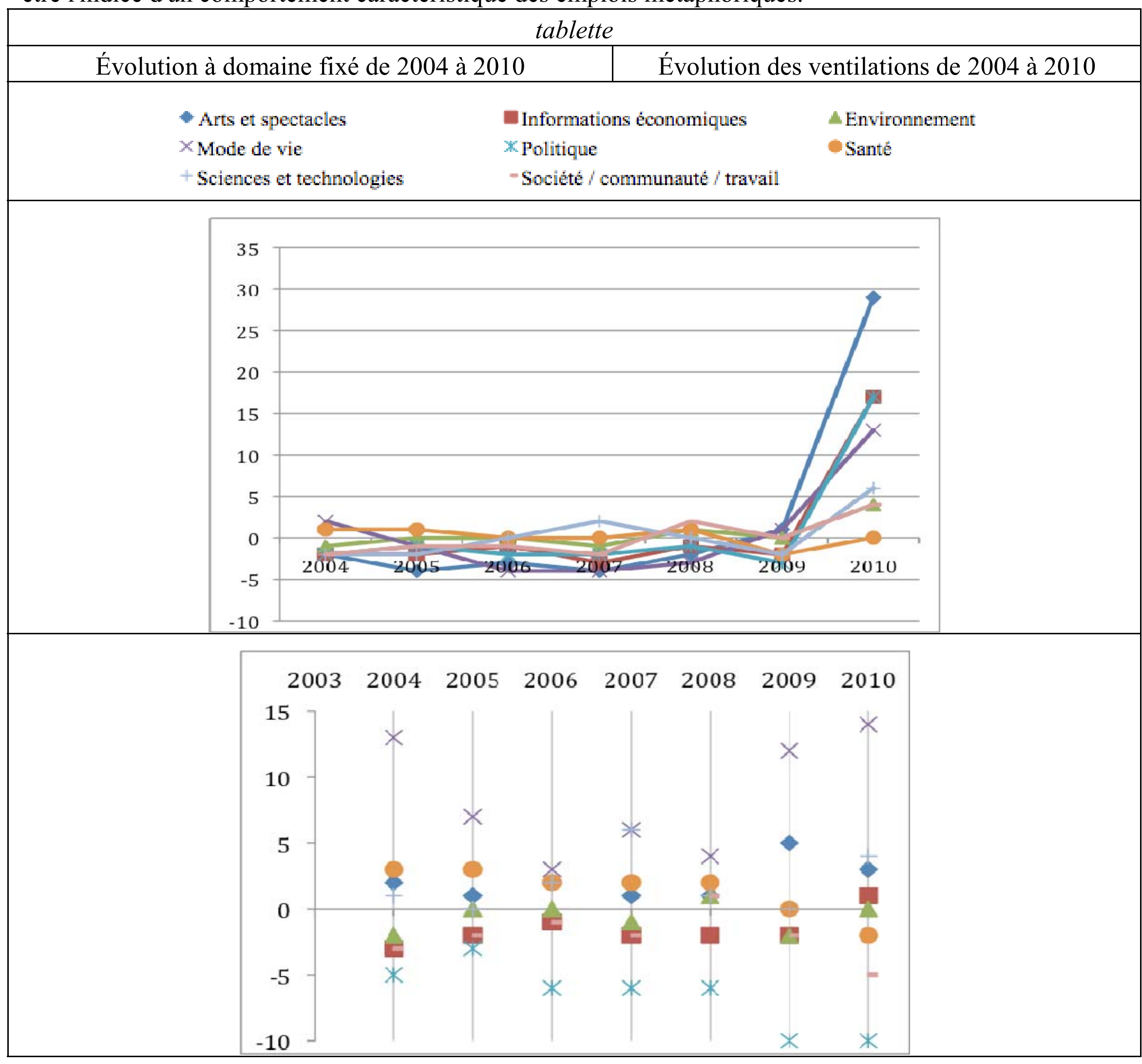


Figure 5d. Évolution de *tablette* à domaine fixé et dans les ventilations de 2004 à 2010

Analyse de la cible *tablette* (figure 5d). À domaine fixé, l'évolution des domaines est relativement stable jusqu'en 2009 et elle connaît un pic en 2010 dans tous les domaines, sauf en Santé. La ventilation des domaines a une configuration similaire d'année en année jusqu'en 2009. À partir de 2009, une

reconfiguration s'amorce, puis se renforce en 2010. SAN s'efface, ECO s'impose (gain de trois rangs, passage d'une sous-représentation à une surreprésentation en 2010), SCI se renforce. L'évolution des domaines témoigne d'une amorce de changement, apparue de façon assez brutale. Le changement pourrait amener une reconfiguration des domaines d'emploi, avec plusieurs nouveaux domaines d'emploi, notamment ECO et SCI. Comme le changement se produit en 2010, le recul n'est pas suffisant pour savoir si l'évolution de sens perdure ou non.

Les étapes précédemment décrites correspondent à différentes façons d'extraire de l'information des empreintes de fréquence thématisées. À l'issue du processus, nous avons obtenu d'abord une liste révisée de cibles lexicales, avec élimination des cibles sans variations domaniales remarquables (*dangereux*, *délétère, moléculaire* et *raz-de-marée*). Pour les cibles retenues, nous disposons d'une première qualification du nouveau sens à partir d'étiquettes de domaines. À chaque cible lexicale sont associés un ou plusieurs domaines dans lesquels la cible s'est diffusée de façon significative. Dans le cas où plusieurs domaines sont présents, des coefficients leur sont affectés pour évaluer l'importance de la diffusion.

6 Articulation du profil domanial en corpus au sens lexicographique

Une fois les informations analysées en corpus, il convient de les articuler à l'information sémantique antérieure, c'est-à-dire aux sens codés dans des ressources lexicographiques. Pour cela, nous utilisons les définitions d'un dictionnaire, le *Trésor de la Langue Française informatisé* (*TLFi* ; Dendien et Pierrel, 2003). L'articulation des informations textuelles et lexicographiques nécessite d'établir des correspondances entre les domaines propres à chaque ressource, puis d'interpréter les changements en corpus relativement aux sens codés.

6.1 Établir les correspondances entre domaines textuels et lexicographiques

Pour mettre en parallèle les informations du corpus avec les sens existants, nous utilisons comme vivier de sens codés une ressource électronique, le *Trésor de la Langue Française informatisé* (*TLFi*).

Les définitions du *TLFi* sont accessibles en ligne, mais elles sont également encodées dans une base de données, SEMEME, dérivée du *TLFi* dans le cadre d'un projet interne. À chaque entrée est associé un fichier XML qui encode la subdivision des entrées associées à des mots-vedettes en définitions. Le balisage précise si la définition est dépendante d'un domaine. Cette base de données est couplée à la plateforme SEMY (Grzesitchak, 2008) qui extrait les étiquettes de domaines associées à chaque entrée.

Par ailleurs, au niveau de la macrostructure, les domaines du *TLFi* s'organisent selon une structure hiérarchique, avec différents niveaux d'imbrication. Cette structure est indirectement accessible à partir du moteur de recherche intégré au *TLFi*. Elle est également encodée dans un fichier établi à partir de la nomenclature originelle des domaines utilisée lors de la rédaction du *TLFi*. Les informations de la macrostructure sont utilisées pour ramener les domaines trop précis des définitions à des domaines de premier niveau, c'est-à-dire des domaines correspondant à des subdivisions principales. Ce sont ces domaines principaux qui sont appariés aux domaines du corpus.

Pour relier les domaines de niveau principal du dictionnaire et ceux du corpus, deux critères sont utilisés :

- **Des analogies de premier niveau entre étiquettes de domaines**. Ainsi, le domaine Arts du TLFi est associé au domaine du corpus Arts et spectacles (ART) ; le domaine Économie du *TLFi* est associé à Informations économiques (ECO) ; le domaine Politique du *TLFi* est associé à Politique et relations internationales (POL).
- **Des analogies de second niveau entre sous-domaines des nomenclatures** ou, dans le cas des domaines du corpus, avec les mots-clés des définitions de chaque domaine[6]. Ainsi, dans le corpus, Mode de vie (MOD) est défini par le mot-clé *loisirs*. Il est associé au domaine Loisirs du *TLFi*. De même, SAN est associé à Biologie.

[6] Dans la base de données Factiva, une note d'information est associée à chaque domaine (ou sujet). Cette note précise le champ que recouvre le domaine. Nous considérons comme mots-clés les noms et adjectifs de ces notes d'information.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100304

Environnement (ENV) est associé à Sciences de la terre et à Biologie (la définition fournie par le corpus précise que les documents de ce domaine abordent des questions de risques pour la santé).

Dans certains cas, ces critères ne suffisent pas et nécessitent des connaissances externes complémentaires. Ainsi, Science et technologie (SCI) est associé aux domaines du *TLFi* Astronomie, Chimie, Information, Mathématiques, Physique et Sciences de la terre. Société (SOC) est associé à Droit, Ethnologie, Philosophie et Religion.

Une fois l'appariement réalisé, les domaines présents dans les définitions des cibles peuvent être reliés aux domaines du corpus (*cf.* figure 6). À ce stade, nous observons seulement les domaines qui apparaissent éventuellement dans les entrées de chaque cible lexicale, sans prendre en compte l'existence de définitions non dépendantes de domaines. Les correspondances sont propres aux sens codés, donc aux anciens sens, indépendamment des sens émergents. Elles résultent des informations des dictionnaires.

	Dictionnaire		Corpus
	Domaines des définitions	Domaines de niveau principal	Domaines correspondants
toxique	Biologie, Chimie, Médecine, Pharmacologie, Physiologie	Biologie, Chimie	SAN, SCI, ENV
tsunami	Sciences de la terre	Sciences de la terre	ENV, SCI
tablette	Histoire	Histoire	Pas de domaine correspondant

Figure 6. Appariement des domaines des définitions aux domaines du corpus

6.2 Analyse du nouveau sens en termes d'enrichissement et de reconfiguration des domaines d'emploi

L'information relative aux anciens sens est articulée à l'information obtenue en corpus. La figure 7 précise le type de nouveau sens en fonction des domaines émergents en corpus, ainsi que les domaines spécifiques et l'existence de définitions non domanialisées (emplois généraux, métaphoriques, etc.).

	Corpus	Dictionnaire	
	Hypothèse sur le nouveau sens en fonction des domaines	Sens spécifiques à certains domaines	Sens généraux (non domanialisés)
toxique	Sens spécialisé en ECO	Sens en ENV, SCI, SAN	Un sens général
tsunami	Un sens général (multidomaines). Il y a une relative instabilité dans l'émergence de domaines.	Sens en ENV, SCI	Pas de sens général
tablette	Un sens général ou plusieurs sens spécifiques en SCI et ECO principalement	Sens dans un domaine absent du corpus	Plusieurs sens généraux

Figure 7. Mise en relation des domaines de l'ancien sens en dictionnaire et du nouveau sens en corpus

La confrontation des deux ressources permet de faire les hypothèses suivantes :

- **Pour *toxique***, il n'y a pas de sens spécialisé correspondant à celui qui émerge en corpus, propre à un seul domaine, ce qui va dans le sens de l'émergence d'un nouveau sens spécialisé en économie.
- **Pour *tsunami***, seul un sens spécialisé existe dans le dictionnaire alors que des domaines multiples émergent en corpus, ce qui invite à voir l'apparition d'un nouveau sens général.
- **Pour *tablette***, plusieurs hypothèses se présentent. Il peut y avoir soit une reconfiguration des emplois (les sens généraux prennent le pas sur les sens domanialisés en histoire), soit l'émergence de plusieurs sens spécifiques, en Science et en économie vraisemblablement, soit l'apparition d'un sens général distinct des sens généraux existants, qui pourrait être plus proche du domaine scientifique ou économique que d'autres domaines.

À ce stade, les informations délivrées par les domaines ne permettent pas de préciser davantage le nouveau sens. L'information générique qu'ils donnent nécessite d'être complétée par des informations plus précises.

7 Conclusion et perspectives

Nous avons proposé une méthodologie outillée pour qualifier la diffusion d'un nouveau sens lexical en corpus. Cette méthodologie a été étayée par des mesures statistiques et par le jeu sur diverses structurations de l'espace textuel, en fonction du temps, de domaines d'emploi et du voisinage de l'unité lexicale ciblée. L'évolution de sens a été décrite à l'aide de descripteurs génériques, les domaines. Ceux-ci ont servi à détecter l'émergence du nouveau sens en discours, à le qualifier et à articuler l'information sémantique provenant des ressources textuelles avec le sens codé dans un dictionnaire. Les étapes successives ont permis d'obtenir des informations de plus en plus précises et structurées. La diffusion temporelle d'une cible lexicale dans de nouveaux domaines a d'abord été analysée à travers le contraste de deux périodes pour montrer l'existence d'une rupture, puis à travers l'analyse plus fine d'une succession de périodes, pertinente pour décrire le processus qu'est la diffusion d'un nouveau sens. Les nouveaux domaines ont été recherchés par recoupement de deux axes d'observation, l'évolution temporelle à domaine fixé et l'évolution de la ventilation des domaines. Enfin, les domaines du corpus susceptibles de qualifier le nouveau sens ont été confrontés aux domaines des définitions lexicographiques des cibles lexicales. Les recoupements avec le sens codé ont permis de faire des hypothèses sur le type d'évolution du sens codé en termes de reconfiguration ou d'enrichissement des définitions.

Le protocole s'est construit grâce aux structures complexes propres aux ressources textuelles et lexicographiques, même si ces structures n'ont été exploitées que partiellement. La structure textuelle a été analysée de façon relativement poussée. La structure du dictionnaire est apparue en terme de macrostructure (organisation globale des domaines) et de microstructure (dépendance des définitions aux domaines). Elle n'a été exploitée que de façon limitée, de même que l'articulation de la structure textuelle à la structure lexicographique. Pour progresser dans l'automatisation des interactions entre corpus et dictionnaire, la mise en œuvre de techniques d'appariement plus approfondies est nécessaire.

Par ailleurs, la qualification du sens que nous avons proposée reste de haut niveau, elle repose sur des descripteurs génériques. Pour évoluer vers une caractérisation plus complète et plus précise du nouveau sens, le protocole pourrait être étendu à d'autres niveaux de description, notamment lexicaux et infra-lexicaux. L'acquisition de contenu sémantique s'appuierait sur une recherche d'information par strates, avec l'analyse successive d'unités de granularité sémantique de plus en plus fine, selon un protocole dont les grandes lignes ont été proposées dans (Reutenauer, 2012). Les domaines joueraient un rôle central et seraient la première étape de qualifications successives du sens, de plus en plus nuancées. Leur analyse servirait de clé de structuration principale de l'information et elle serait réutilisée pour sélectionner ou structurer les unités plus précises, par exemple des unités lexicales ou, à un niveau encore plus fin, des traits sémantiques.

Références bibliographiques

Bauer, L. & Renouf, A. (2000). Contextual clues to Word-Meaning. *International Journal of Corpus Linguistics, 5-2*, 231-259.

Béjoint, H. (2009). Lexicographie et linguistique : le domaine anglais. *Lexique, 19*, 117-158.

Brunet E. (2007). Fréquences et séquences : mise en œuvre dans Hyperbase. *Lexicometrica, 7.*

Cabré, M., Domenech, M., Estopa, R., Freixa, J. & Sole, E. (2003). L'Observatoire de néologie : conception, méthodologie, résultats et nouveaux travaux. *L'innovation lexicale*. Paris : Honoré Champion, 125-147.

Cusin-Berche, F. (2003). *Les mots et leurs contextes*. Paris : Presses Sorbonne Nouvelle.

Dendien, J. & Pierrel, JM. (2003). Le Trésor de la Langue Française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence. *TAL, 44-2*, 11-37.

Grzesitchak, M. (2008). Annotation sémantique : profilage textuel et lexical. *Lexicographie et informatique : bilan et perspective, colloque à l'occasion du 50e anniversaire du projet du Trésor de la Langue Française*. Nancy.

Harris, Z. (1968). *Mathematical Structures of Languages*. New York : John Wiley and Sons.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100304

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. *Euralex*, 105-116.

Lafon, P. (1984). *Dépouillements et statistiques en lexicométrie*. Genève-Paris :Slatkine-Champion.

Lemnitzer, L. & Ule, T. (2011). *Die Wortwarte - auf der Suche nach den Neuwoertern von morgen*. Rapport technique, Berlin-Brandenburgische Akademie der Wissenschaften. URL : http ://www.wortwarte.de/.

Perlerin, V. (2004). *Sémantique légère pour le document. Assistance personnalisée pour l'accès au document et l'exploration de son contenu.* Thèse de doctorat, Université de Caen, Basse-Normandie.

Picton, A. (2009). *Diachronie en langue de spécialité. Définition d'une méthode linguistique outillée pour repérer l'évolution des connaissances en corpus. Un exemple appliqué au domaine spatial*. Thèse de doctorat, Toulouse 2.

Rastier, F. (1987). *Sémantique interprétative*. Paris : PUF.

Rastier, F., Valette M. (2009). De la polysémie à la néosémie. *Le français moderne*, *77*, 97-116.

Rayson, P., Archer, D., Piao, S. & McEnery T. (2004). The UCREL semantic analysis system, Workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks. *LREC 2004*, 7-12.

Reutenauer, C. (2012). *Vers un traitement automatique de la néosémie : approche textuelle et statistique.* Thèse de doctorat, Université de Lorraine.

Rey, A. (2008). *De l'artisanat des dictionnaires à une science du mot. Images et modèles*. Paris : A. Colin.

Sablayrolles, J. (2000). *La néologie en français contemporain : examen du concept et analyse de productions néologiques récentes.* Paris : Champion.

Salem, A., Lamalle, C., Martinez, W., Fleury, S., Fracchiolla, B., Kuncova, A. & Maisondieu, A. (2003). *Lexico3 - Outils de statistique textuelle. Manuel d'utilisation*, Syled-CLA2T, Université de la Sorbonne nouvelle - Paris 3.

Victorri, B. & Fuchs C. (1996). *La polysémie, construction dynamique du sens*. Paris : Hermès.

Viprey, J. & Schepens, P. (2010). Dérivation lexicale et dérive du discours : mutualiser, mutualisation. In : *10th International Conference Journées d'Analyse Statistique des Données Textuelles (JADT 2010)*, 489-498.